MorphAdorner

Morphological Adornment of English Language Literary Texts

Corpora Space Workshop II
June 7, 2011

Why "Adornment"?

- Terms like annotation, tagging, etc. have too many alternate and confusing meanings
- Adornment harkens back to medieval sense of manuscript adornment or illumination -- attaching pictures, marginal comments, etc. to texts
- Morphological adornment is thus the process of "adorning" words with morphological information such as part of speech, lemma, standardized spelling, semantic category, etc.

Sample Adornment Processes

- Tokenization
- Sentence Boundary Recognition
- Spelling Normalization
- Part of Speech Tagging
- Lemmatization
- Name Extraction

MorphAdorner Pipeline

- MorphAdorner provides "skeleton" for pipelining adornment processes
- Use of Java interfaces for adornment processes allows easy substitution of different implementations into pipeline (e.g., Template method pattern)
- Straightforward to wrap adornment processes as web services using Rest-like interfaces

MorphAdorner Audience

- MorphAdorner intended as a programmer's construction kit, not an end-user program
- MorphAdorner can be used to create customized end-user programs for morphological adornment
- Released to public under open source license in April 2009
- Some continued updates to training data during 2010
- Initial work on wrapping MorphAdorner facilities as RESTful web services in May 2011

Sample web service example: Lemmatizer

Find lemma form of early modern English spelling "strykynge"

http://localhost:8182/lemmatizer?
spelling=strykynge&standardize=true&wordClass=verb&wordClass2
=&corpusConfig=eme

Lemmatizer example (cont.)

XML result:

```
<LemmatizerResult>
<spelling>strykynge</spelling>
<standardSpelling>striking</standardSpelling>
<corpusConfig>eme</corpusConfig>
<wordClass>verb</wordClass>
<wordClass2/>
<lemma>strike</lemma>
<standardize>true</standardize>
</LemmatizerResult>
```

Lemmatizer example (cont.)

JSON result:

```
{"spelling":"strykynge",
"standardSpelling":"striking",
"corpusConfig":"eme",
"wordClass":"verb",
"wordClass2":"",
"lemma":"strike",
"standardize":true}
```

Pos Tagging Example

Text to adorn: Mary had a little lamb.

http://localhost:8182/partofspeechtagger?text=Mary+had+a+little+lamb.&corpusConfig=ncf

(In practice we would use HTTP post to allow for long texts.)

Pos Tagging Example (cont.)

```
<PartOfSpeechTaggerResult>
<text>Mary had a little lamb.</text>
<lexicon/>
–<sentences>
–<list>
<string>Mary</string>
<string>had</string>
<string>a</string>
<string>little</string>
<string>lamb</string>
<string>.</string>
</list>
</sentences>
–
```

Pos Tagging Example (cont.)

```
–<taggedSentences>
–<list>
–<AdornedWord>
<token>Mary</token>
<spelling>Mary</spelling>
<standardSpelling>Mary</standardSpelling>
<lemmata>Mary</lemmata>
<partsOfSpeech>np1</partsOfSpeech>
<tokenType>0</tokenType>
</AdornedWord>
   …
```

Other MorphAdorner Facilities

- Language Recognition
- Name Standardization
- Parser
- Pluralizer
- Statistics (Dunning's Log Likelihood and others)
- Stemming
- Syllabification
- Text Segmenter
- Text Summarization
- Thesaurus (synonyms and antonyms)
- Verb Conjugator

Other hooks

- Custom output adapters for generating input to Xaira, the Corpus Workbench (CWB), Lucene, and word lists in a variety of formats.
- MorphAdorner can also be integrated with Gate (and therefore UIMA).

Personal Goal: Comprehensive Lexicon

- Spelling and variants with date information
- Frequencies of occurrence across centuries
- Frequencies of occurrence for specific genres
- Frequencies of occurrences by part of speech
- Lemmata by part of speech
- Allows morphological adornment processes to use a standardized lexicon ID

MorphAdorner and Project Bamboo

- Designed to work with several of the corpora already designated for use in the initial phase of Project Bamboo
- More aware of potential problems and pitfalls than other existing software for adornment of texts
- NUPos tag set allows adornment of English texts from Middle English to present (diachronic corpora)
- Licensed under a very non-restrictive NCSA style open source license

Summary

- MorphAdorner started providing basic morphological adornment in December 2006
- Used in WordHoard, Monk, and Virtual Orthographic Normalization projects
- First public release in April 2009 under NCSA style open source license
- Updated training data during 2010 and 2011
- Added initial RESTful interfaces in May 2011
- Looking to integrate with existing/forthcoming web services from Project Bamboo and others
- Future work subject to change depending upon grant support and other project workload demands